Online Interactive Data Analysis of Multi-Sensor Data Using Giovanni

Architectural Growing Pains

Stephen Berrick, NASA GSFC

December 7, 2005

Acknowledgments and Resources

- Acknowledgments:
  - NASA:
    - Greg Leptoukh
  - SSAI:
    - Hualan Rui (lead), Jim Acker, Suraiya Ahmad, Tim Dorman, John Farley, Irina Gerasimov, Arun Gopalan, James Johnson, Jason Li, Stephen Maher, Jianping Mao, Andrey Savtchenko, Bill Teng, Xiaoping Zhang, Tong Zhu
  - GMU:
    - Zhong Liu, Suhung Shen
  - Funding:
    - Yoram Kaufman, several REASoN CANs

- Resources:
  - http://giovanni.gsfc.nasa.gov/

What is Giovanni?

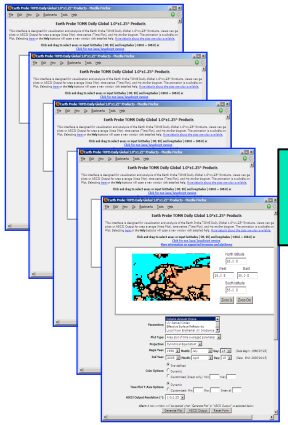
- Giovanni is the **G**ES-DISC (Goddard Earth Sciences Data and Information Services Center) **I**nteractive **O**nline **V**isualization **AN**d a**N**alysis **I**nfrastructure.
- Giovanni was conceived as a research tool to increase the usability of earth science data sets.
- Giovanni features a simple Web user interface allowing rapid access to data, establishment of spatial and temporal criteria, and a variety of output options.
- The data analysis engine provides rapid statistical analyses and generation of area average plots, time plots, Hovmoller latitude vs. time and longitude vs. time plots, vertical profiles, data set intercomparisons, and anomaly analysis.
- There are currently 8 Giovanni instances, each targeted to a specific science community.

Giovanni

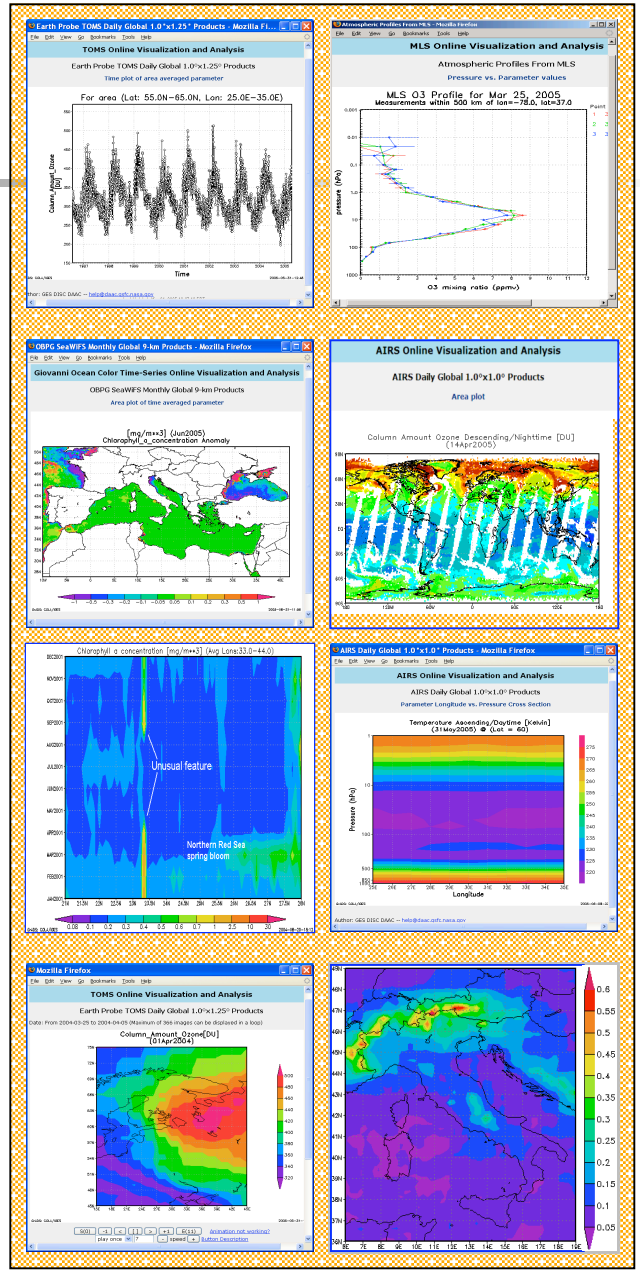
Data Inputs

- MLS Aura
- OMI Aura
- MODIS Aqua
- AIRS Aqua
- MODIS Terra
- SeaWiFS
- TRMM
- HALOE UARS
- TOMS EP Nimbus

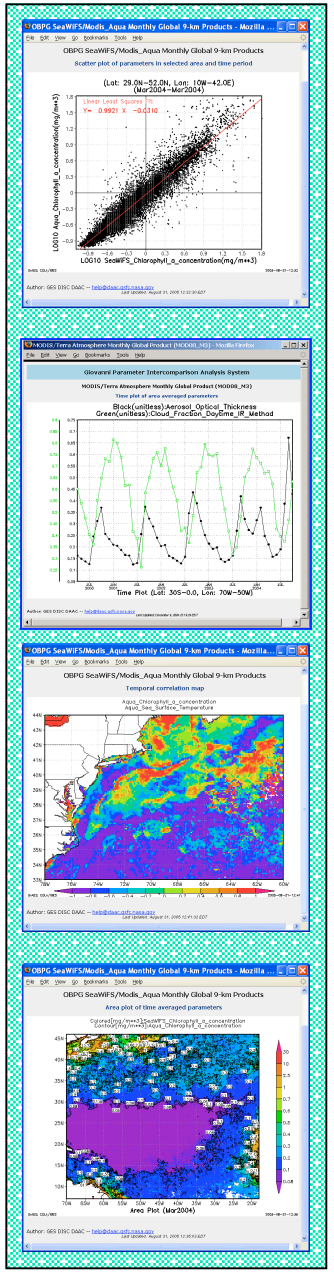
Giovanni Instances



Single Parameter View



Parameter Intercomparison



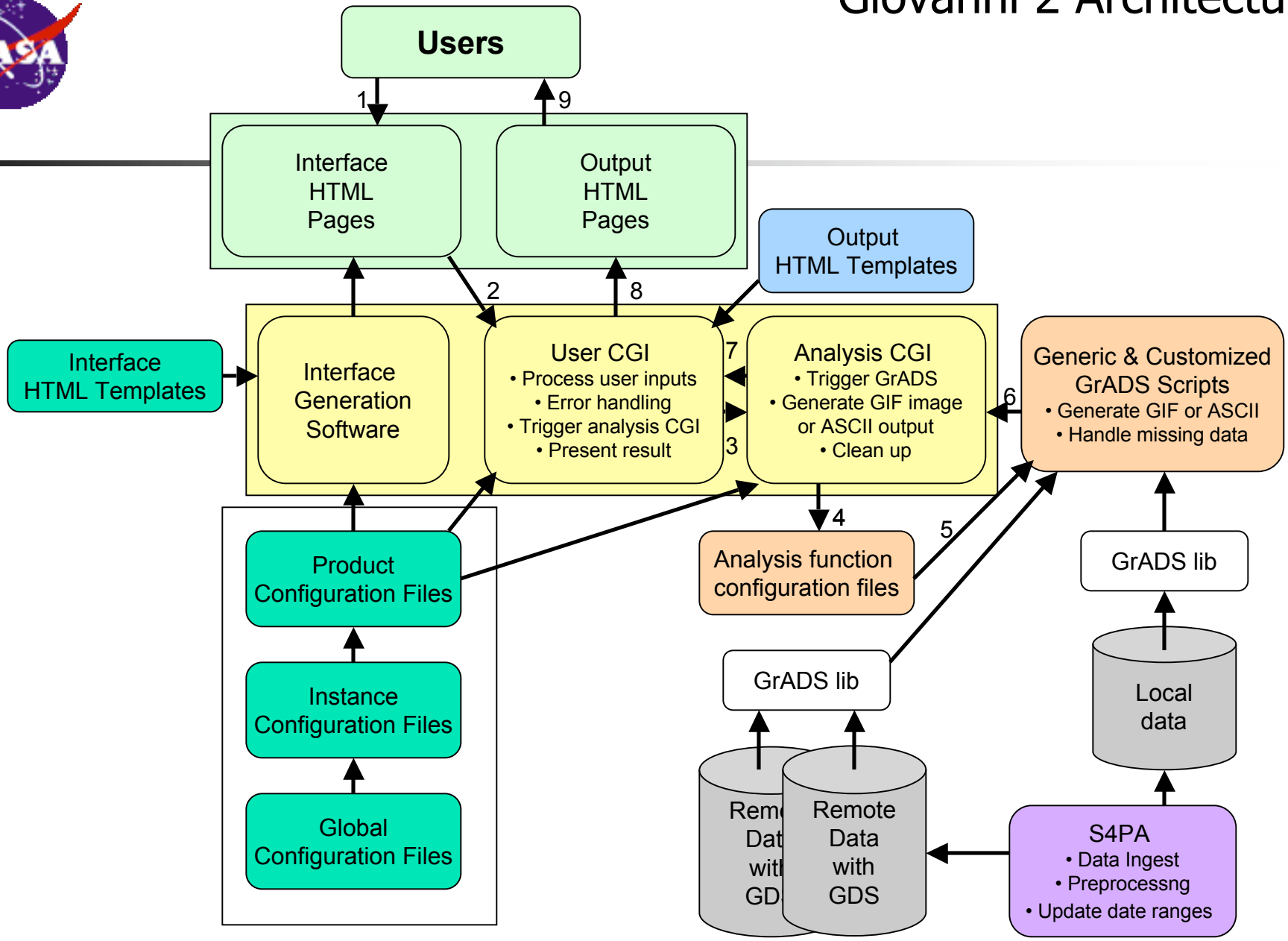
Current Giovanni Instances

- Agricultural (http://agdisc.gsfc.nasa.gov/Giovanni/)
- AIRS (http://reason.gsfc.nasa.gov/Giovanni_airs3d/)
- Aura MLS (http://reason.gsfc.nasa.gov/Giovanni_mls3d/)
- MODIS Aerosols (http://aerodisc.gsfc.nasa.gov/Giovanni/movas/)
- Ocean Color Time-Series Project (http://reason.gsfc.nasa.gov/Giovanni/)
- TOMS and Aura OMI (http://reason.gsfc.nasa.gov/Giovanni_toms/)
- TRMM (http://lake.nascom.nasa.gov/tovas/)
- UARS/HALOE (http://reason.gsfc.nasa.gov/Giovanni_haloe3d/)
- Or, go to http://giovanni.gsfc.nasa.gov/ to see them all
- And more on the way…

Giovanni 2 Architecture

- CGI scripts in Perl
- Perl configuration files
- HTML templates
- Read software and graphics generation in GrADS
- GrADS Data Server (GDS)

Giovanni 2 Architecture



Architectural Issues

- Data location transparency
- Data visualization
- Data formats
- Data regridding
- User interface
- Browser/platform support
- GrADS issues
- Documentation
- Support for algorithms

Data Location Transparency

Solution	Advantages	Disadvantages
Data Replication:	Disk is cheap and most data sets are small	Data management more difficult
	Quick response since data is local	Not well suited for interoperability
GrADS Data Server (GDS):	Allows for fairly sophisticated server side subsetting/averaging	Not well supported
	Based on OPeNDAP or some fork of it	Client must be in GrADS
OPeNDAP:	Widely supported, lots of client support	Less sophisticated subsetting/averaging than GDS
	Supports interoperability	Does not support HDF5 (but should soon)
	Works with most of our data formats	

Data Visualization

- Missing data
  - How to represent? How not to mislead?
- Data quality
  - Often a case of higher quality sparse data vs. good coverage at lower quality. Should allow user to decide.
- Representing errors
  - In original data
  - Due to Giovanni averaging, regridding, and analysis
- Many ways to represent data – Which is best?
  - Often depends on target community

Data Formats

- Mostly HDF4, but a lot of HDF5 too, but little uniformity
- GrADS cannot read HDF5
- Data are preprocessed into binary files

Data Gridding/Regridding

- Regridding necessary when doing multi-sensor parameter intercomparisons
- Gridding to Level 3 from Level 2 data
  - Custom, on-demand Level 3 products

User Interface

- General
  - No single interface applies to all Giovanni supported data or communities
  - Giving user too many options leads to hard to use interface
- Spatial selection applet
  - Java applet, not well supported on all browsers
- Section 508 and other compliance
  - Must support non-Java, non-JavaScript users
  - Changing requirements
  - Tightening of requirements

Browser/Platform Support

- Giovanni needs to support most browsers and platforms of our target audience
  - MS Windows with IE, Firefox, Netscape, or Opera
  - Mac OS X with Safari, Firefox, Netscape, or IE
  - Linux/UNIX with Firefox, Netscape, or Opera
  - Lynx command line browser (Section 508)
- Testing more difficult
- Need to support users who turn off Java or JavaScript (Section 508).

GrADS Issues

- Works only for gridded or station data
- Has limitations on data file names
- Has limitations on where scaling factors are located, also only support linear scaling
- Limited handling of multiple fill values
- Limited flexibility on plots and images (as compared with IDL, for example)
- Doesn’t support HDF5

Documentation

- Documentation has 4 components:
  - How to use Giovanni
  - Information about the original data
  - What Giovanni does to the original data
  - Information about any subsetted data available for download
- Good documentation is time consuming, but also critically important
- User needs to be able to easily find information
- User needs to be aware of the appropriateness of the data for a particular application or use

Support for Algorithms

- Many algorithms for reading and doing analysis already exist and they don’t conform to any standard. How best to reuse this code?

Giovanni V3 Requirements

General	Interface	Algorithms	Data	Graphics
Current V2 capabilities	Component based	Decoupled, standard interface	OPeNDAP, Local, GDS, S4PA, Web services	Decoupled
Multi-parameter intercomparisons	Recipes	Chaining	Minimize preprocessing, replication	GrADS, IDL, Matlab, others
Distributed	Documentation	Language agnostic	Common format in Giovanni	
Asynchronous support (E-mail, RSS)	Browser/platform support	Encourage modularity and reuse	Subsetted data only	
Reasonable defaults with override capability	Section 508 & other NASA/Federal		Data fusion or analysis	
Provide metrics	User feedback			
"Our" and "My" Giovanni				

Giovanni 3 Architecture: What we know

- Algorithms via standard Web Services interface
- Some common data format
- Workflow management
- XML based
- YAGNI ("You aren't going to need it!") development approach

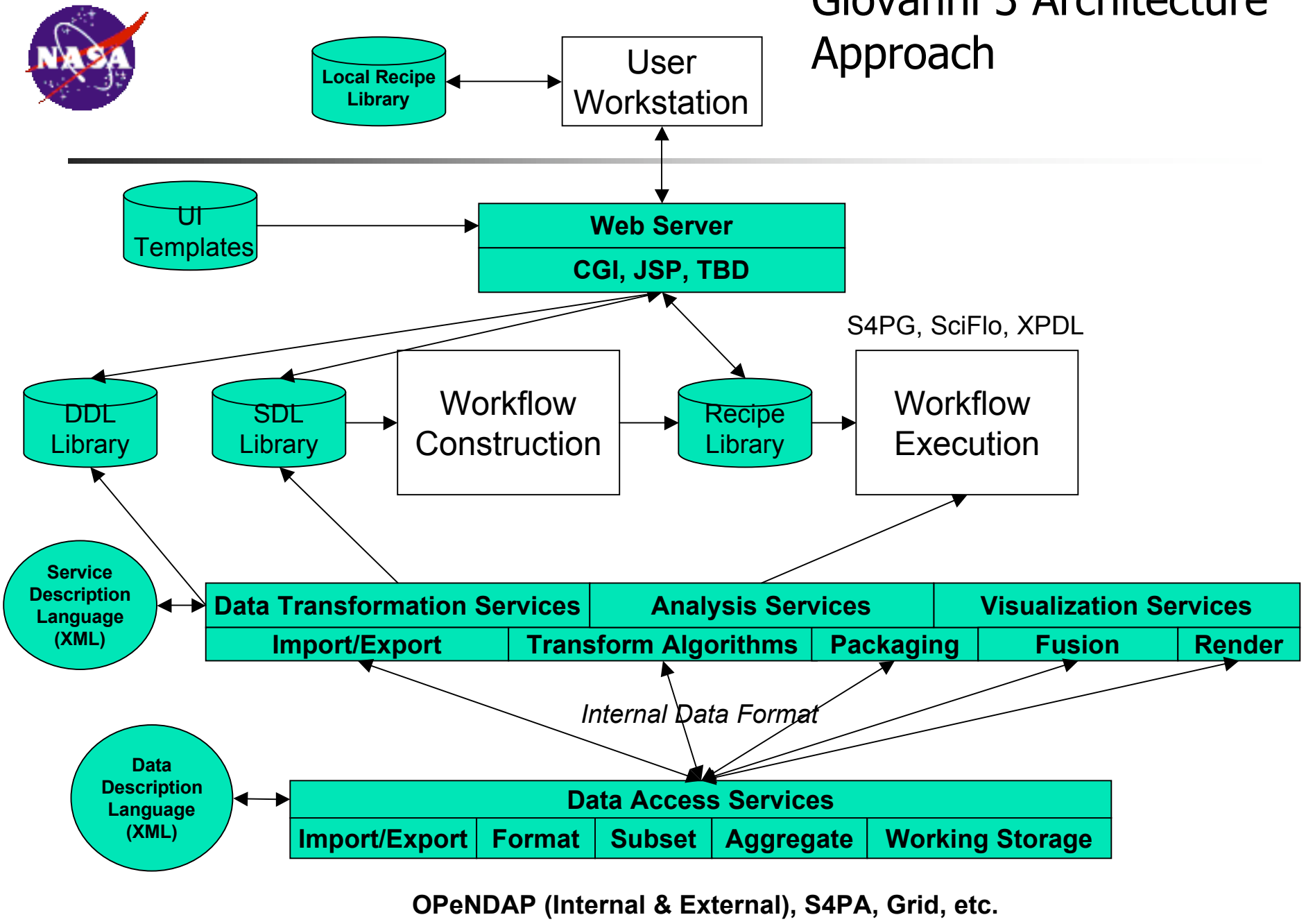
Giovanni 3 Architecture: Trades

- What is our common data format?
  - NetCDF, HDF, some other?
- User Interface:
  - Perl CGI, Java Server Pages (JSP), Mason?
- XML Language
- Algorithm design and modularity
- Dealing with asynchronous nature
- Workflow management and engine:
  - S4P/S4PM based?
  - GENESIS SciFlo, Kepler?
  - Many others?

What is S4P/S4PM?

- S4PM is "Simple, Scalable, Script-Based Science Processor for Measurements"
  - System for highly automated processing of science data. It is the main processing engine at the GES DISC. In addition to being scalable up to large processing systems, it is also scalable down to small, special-purpose processing strings.
  - S4PM built upon S4P core (as is S4PA)
  - http://disc.gsfc.nasa.gov/techlab/s4pm/
- Supports a workflow engine where workflows are implicit and described through production rules.
- Perl based
- Developed using Extreme Programming/Scrum approach
- Used reliably and cheaply for over 4 years (TRL 9)
- Released under NASA Open Source Agreement
- http://sourceforge.net/projects/s4pm
- BUT: S4P/S4PM would need non-trivial modifications to be able to support Giovanni (S4PG?)

Giovanni 3 Architecture Approach



OPeNDAP (Internal & External), S4PA, Grid, etc.